# Character-level Convolutional Neural Networks を用いた新聞社間の記事の違いの解析の試み

宗里 駿　　小谷 龍ノ介　　彌冨 仁

法政大学 理工学部 応用情報工学科
{hayao.munesato.7w@stu., iyatomi@}hosei.ac.jp

**概 要**

客観性、正確性、公平性が高いことが望まれる新聞やテレビといったマスメディアにおいても発信者の意図による報道の偏りが存在する。本報告ではメディアの報道の違いを客観的に捉え、将来的にメディアの偏りの定量化を行うための前段階として、新聞社の各記事を character-level convolutional neural networks (CLCNN) で解析し、各記事の発行会社を推定すると共に、その判断のための重要な手掛かりとなった部位、つまりその新聞社の特徴的な表現と考えられる部位の可視化を行った。本実験では、Web で公開されている大手新聞社 4 社の政治、経済、国際面計 19,442 記事の内、記事の長さが 350 文字以上である新聞記事計 4,836 記事を対象に解析した。その結果、新聞記事からの発行会社の推定問題において 4-fold cross validation で評価したところ平均 90.0%の識別精度を実現した。新聞社の特徴を表していると考えられる可視化された部位は、記事により多様であり今後の解析が必要であるが、新聞社によっては人名や固有名詞に強く反応が現れる傾向が見られた。

## 1 背景

様々なメディアが発信する膨大な情報を容易に得ることが可能になった今日、正確な情報を得るためには情報の客観性・正確性・公平性を考慮した的確な判断が求められている。新聞は正確で公正な記事と責任ある論評により社会の要望にこたえ、公共的、文化的使命を果たすことが責務であると定められている [1]。しかしながらこうしたマスメディアにおいても発信者の意図による報道の偏りが存在する。河野らは、テレビ局各局の報道の分析を行い、ニュースのトピック間に報道量の差やイメージの差があり、その差異はテレビ局によって異なることを示した [2]。本研究では、こうした違いの客観的な定量化や可視化を行うための基礎検討として、特に客観的な報道が求められる大手新聞社 4 社の新聞社を対象に、記事内容から、それらの違いについて客観的に得られる違いについて検討を行った。

自然言語処理において、文字列を必要に応じた前処理を経て時系列 $n$ をベクトル表現に変換したのち、それらを何らかのモデルで処理を行うことが一般的である。recurrent neural networks (RNN) は非線形な時系列データを単純なモデルで効率よく学習できるため、これまで広く使われてきた。しかしながら、RNN は確率的勾配降下法によりパラメータの最適化を実施しているため、時間方向の勾配が消失してしまう問題 [3] が存在し、長い文字列の扱いに困難が存在していた。その問題に対応するために入力ゲート、出力ゲート及び忘却ゲートを用いた long short-term memory (LSTM)[4] が提案され、長期依存性のある時系列データの学習が可能となった。また、深層学習技術の最も中心的なモデルとして知られ、主に画像認識タスクに用いられる convolutional neural networks (CNN) の入力を 1 次元にして文書分類タスクに応用した例も見られ高い精度を上げている [5]。

一方で日本語の解析は、英語と異なり困難な形態素解析が必要であり、また文字の種類が極めて多いため、主に英語で利用される手法をそのまま利用できない。島田ら [6] は、文章中の各文字を画像とみなし、convolutional auto encoder にて文字形状に基づいた文字埋め込みを行った後、1 次元 CNN である character-level convolutional neural networks (CLCNN) で学習することにより、こうした形態素解析や、文字種類の多様性による困難さを大幅に緩和して、文書識別問題において良好な結果を達成している。

新聞記事の解析について、英文に対しては様々な研究が行われており、新聞社間のメディアの論調の違い

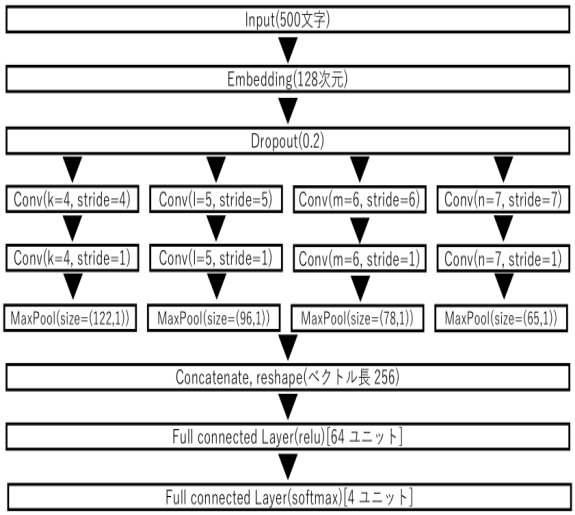


図 1: CLCNN のモデル

表 1: 新聞社別の記事数

| 読売新聞 | 朝日新聞 | 毎日新聞 | 産経新聞 |
|---|---|---|---|
| 718 | 1154 | 1616 | 1348 |

を定量的に解析した研究も見られる。Fortuny ら [7] は、ベルギーの新聞を解析し、各メディアが特定の政党への投票を促すような記載をしていることについて解析を行っている。日本語の新聞記事の内容を対象とした自然言語処理の研究には、見出しの提案や要約、特定のキーワードに関する報道の変化 [8]、特定の傾向を持つ記事の抽出 [9]、経済動向の予測 [10] などがあげられるが、メディア間の違いを定量的に解析する研究は数少ない。

本研究では CLCNN を用いて、各社の新聞記事を客観的に解析し、記事のみから新聞社の推定が可能であることを示したうえで、その特徴的な表現を可視化することで、メディアの報道内容、姿勢の差を今後定量的に評価するための基礎的な足掛かりにする。

# 2 提案手法

本研究では大手新聞社 4 社 (読売、朝日、毎日、産経) の政治、経済、国際面で 2014 年 7 月 23 日から 2015 年 11 月 19 日及び 2017 年 10 月 12 日から 2017 年 12 月 24 日までに取得した計 19,442 記事の内、記事の長さが 350 文字以上計 4,836 記事を対象に解析した。解析した記事数の分布を表 1 に示す。

## 2.1 新聞記事から発行会社の推定

新聞記事の各文字を unicode を用いて符号化を行い、CLCNN を用いて学習を行った。用いた CLCNN のモデルを図 1 に示す。

我々が用いた CLCNN モデルには、500 文字ごとに 1 文字ごとスライドさせながら記事を入力させる。記事が 500 文字に満たない場合には 0 パディングを行い、500 文字を超える場合には、その記事の先頭から 500 文字を用いた。また前述の通り、まず各文字を unicode で 128 次元のベクトルに変換する。続いて、入力文字ベクトルを任意の確率で dropout させる wildcard training を導入した。これは島田ら [6] が CLCNN の過学習を抑えるために提案した手法で、各種文書分類問題で 10%以上の精度向上を実現した優れた効果が報告された手法である。本モデルは、4〜7 の異なるカーネルサイズでの浅い畳み込みを行ったのち、それに続く極めて大きい max pooling 処理を行っている。前者については、Saxe ら [11] が自然言語処理においても、異なるカーネルサイズを用いることで成果を上げた手法を参考に予備実験によりカーネルサイズを決定した。後者については、Raff ら [12] が、長い解析文字列の中から特徴的な部位を検出する際に効果的であったモデルを参考にしている。

CLCNN を用いることで、適切な処理が困難な形態素解析を行わずに処理が可能となる。記事の発行会社の推定については、4-fold cross validation により評価を行った。

## 2.2 新聞社の記事特徴の提示

記事の中で、その新聞社の特徴を表現している部位の可視化を目的に、CLCNN の出力の根拠の提示を、画像中の関心領域の抽出として提案された Zeiler らの手法 [13] を参考に試みた。具体的には、解析対象の CLCNN 出力と、解析対象の文字列内任意の $N$ 文字を null 文字に置き換えた際の CLCNN の出力の差を、その $N$ 文字が識別に与える影響度として記録した。その部位をラスタスキャンさせて加算することにより各記事において、ヒートマップを作成した。これは、どの部位の表現が記事の識別に重要であったか、つまりその新聞社の特徴的な表現であることが期待できる。予備実験の結果、本実験では $N$=5 とした。

表 2: 新聞社推定の confusion matrix

| | predict label(%) | | | | |
|---|---|---|---|---|---|
| | | 読売新聞 | 朝日新聞 | 毎日新聞 | 産経新聞 |
| true label | 読売新聞 | **87.0** | 7.8 | 1.4 | 3.8 |
| | 朝日新聞 | 4.9 | **87.0** | 1.0 | 6.8 |
| | 毎日新聞 | 0.6 | 1.4 | **97.0** | 1.5 |
| | 産経新聞 | 4.7 | 0.6 | 0.1 | **89.0** |

# 3 実験結果

## 3.1 新聞記事から発行会社の推定

記事からの発行会社の結果についての confusion matrix を表 2 に示した。データセット全体での識別精度は平均 90.0%であった。以上の結果より CLCNN を用いることで形態素解析を必要とせずに、新聞記事のみからその発行社を推定するタスクを実現できたといえる。またこのことは、CLCNN が記事の中から新聞社を特徴づける何等かの特徴が存在し、それを獲得できたと考えられる。

## 3.2 新聞社の記事特徴の提示

読売、朝日、毎日、産経の新聞記事に対する発行会社推定において各新聞社の特徴を表すと考えられる表現のヒートマップの例を図 2〜図 5 に示す。CLCNN を用いた新聞記事の発行社推定で高い精度の実現ができていることから、得られた文の発火した部位は、各新聞の特徴を表していると考えられ、概ね単語ごとに発火していることが確認できる。現時点では得られた結果の定性的な解釈のみであるが、読売新聞では比較的人名の発火はあまりせず、朝日新聞は反対に人名に対してよく発火する傾向にあった。毎日新聞はほぼ全ての部位で強く発火する傾向にあった。このことは記事全体が毎日新聞を特徴づける、言い換えれば今回の解析において、毎日新聞の記事は他の 3 誌に比べて表現上の強い特徴が見られにくいとも考えられる。いずれにしても今後、より深く定量的な解析が必要となる。

# 4 おわりに

各新聞社とも固有名詞には強く発火する傾向があったが、同一単語でも新聞社によって発火の差異があった。しかしながら、本研究では定量評価は未だ出来ておらず、今後の課題となる。

# 参考文献


[1] 日本新聞協会, **新聞倫理綱領**, 2006.

[2] T. Kohono, “A comparative content analysis on the tv news coverage in the 1993 and 1996 general election,” *Japanese Journal of Element Studies*, vol. 13, pp. 78–88, 1998.

[3] Y. Bengio, P. Simard, and P. Franconi, “Learning long-term dependencies with gradient descent is difficult,” *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[4] S. Hochreiter and J. Schmidhuber, “Lont short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] Y. Kim, “Convolutional neural networks for sentence classification,” *Proceccing of Conference on Empirical Methods in Neural Information Processing Systems (NIPS)*, pp. 649–657, 2015.

[6] D. Shimada, R. Kotani, and H. Iyatomi, “Document classification through image-based character embedding and wildcard training,” *2016 IEEE International Conference on Big Data*, pp. 3922–3927, 2016.

[7] E.J.Fourtuny, T.D.Smedt, D.Martens, and W.Daelemans, “Media coverage in times of political crisis: A text mining approach,” *Expert Systems Witj Applications*, vol. 39, no. 14, pp. 11 616–11 622, 2012.

[8] S.Koike, S.Yamaguchi, and Y. et al, “Effect of name change of schizophrenia on mass media between 1985 and 2013 in japan: A text data mining analysis,” *Schizophrenia Bulletin*, vol. 42, no. 3, pp. 552–559, 2016.

[9] 興梠紗和, 木村昭悟, 藤代裕之, and 西川仁, “Sns 上での拡散を誘発する web ニュース説明文の調査と自動選択,” **電子情報通信学会論文誌** *D*, vol. J99-D, no. 4, pp. 403–414, 2016.

[10] 吉原輝, 藤川和樹, 関和広, and 上原邦昭, “深層学習による経済指標動向推定,” in *The 28th Annual Conference of the Japanese Society for Articial Intelligence*. Os–24, 2014.

[11] J. Saxe and K. Berlin, “A character-level convolutional neural network with embeddings for detecting malicious urls, file paths and registry keys,” *CoRR arXiv:1710.09435*, 2017.

[12] E. Roff, J. Barker, J. Sylvester, R. Barndon, B. Catanzaro, and C. Nicholas, “Malware detection by eating a whole exe,” *CoRR arXiv:1702.08568*, 2017.

[13] M. D. Zeiler and R. Fergus, “Visualising and understanding convolutional networks,” *CoRR arXiv:1311.2901*, 2013.

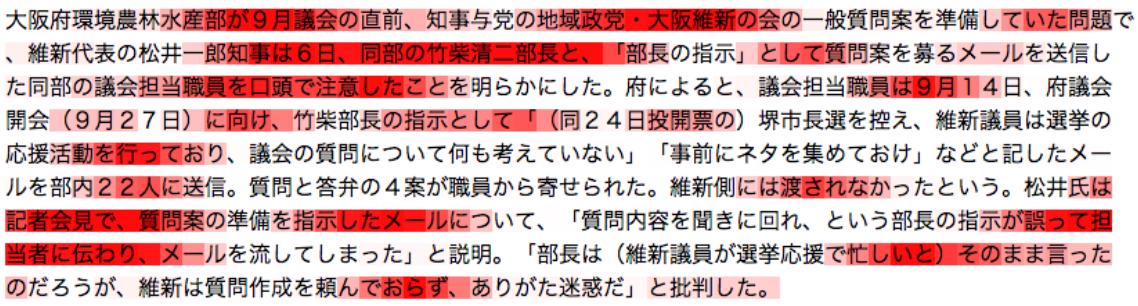
大阪府環境農林水産部が９月議会の直前、知事与党の地域政党・大阪維新の会の一般質問案を準備していた問題で、維新代表の松井一郎知事は６日、同部の竹柴清二部長と、「部長の指示」として質問案を募るメールを送信した同部の議会担当職員を口頭で注意したことを明らかにした。府によると、議会担当職員は９月１４日、府議会開会（９月２７日）に向け、竹柴部長の指示として「（同２４日投開票の）堺市長選を控え、維新議員は選挙の応援活動を行っており、議会の質問について何も考えていない」「事前にネタを集めておけ」などと記したメールを部内２２人に送信。質問と答弁の４案が職員から寄せられた。維新側には渡されなかったという。松井氏は記者会見で、質問案の準備を指示したメールについて、「質問内容を聞きに回れ、という部長の指示が誤って担当者に伝わり、メールを流してしまった」と説明。「部長は（維新議員が選挙応援で忙しいと）そのまま言ったのだろうが、維新は質問作成を頼んでおらず、ありがた迷惑だ」と批判した。

図 2: 読売新聞の特徴を表すと考えられる表現のヒートマップの例

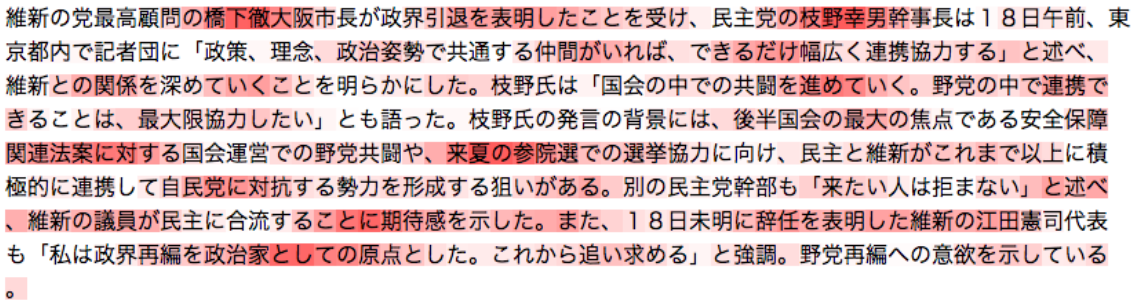
維新の党最高顧問の橋下徹大阪市長が政界引退を表明したことを受け、民主党の枝野幸男幹事長は１８日午前、東京都内で記者団に「政策、理念、政治姿勢で共通する仲間がいれば、できるだけ幅広く連携協力する」と述べ、維新との関係を深めていくことを明らかにした。枝野氏は「国会の中での共闘を進めていく。野党の中で連携できることは、最大限協力したい」とも語った。枝野氏の発言の背景には、後半国会の最大の焦点である安全保障関連法案に対する国会運営での野党共闘や、来夏の参院選での選挙協力に向け、民主と維新がこれまで以上に積極的に連携して自民党に対抗する勢力を形成する狙いがある。別の民主党幹部も「来たい人は拒まない」と述べ、維新の議員が民主に合流することに期待感を示した。また、１８日未明に辞任を表明した維新の江田憲司代表も「私は政界再編を政治家としての原点とした。これから追い求める」と強調。野党再編への意欲を示している。

図 3: 朝日新聞の特徴を表すと考えられる表現のヒートマップの例

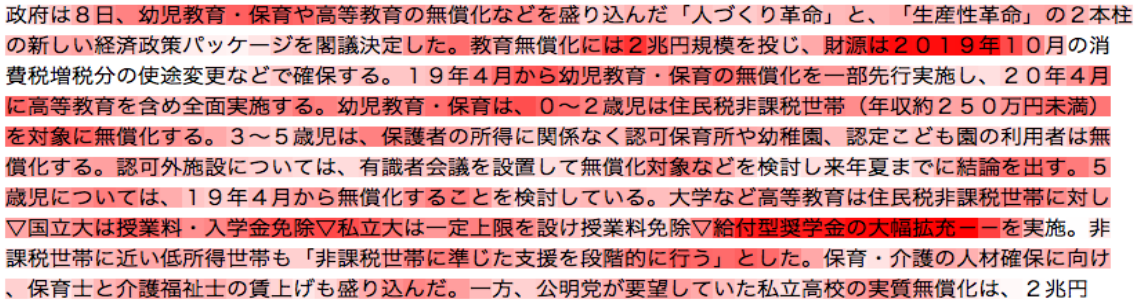
政府は８日、幼児教育・保育や高等教育の無償化などを盛り込んだ「人づくり革命」と、「生産性革命」の２本柱の新しい経済政策パッケージを閣議決定した。教育無償化には２兆円規模を投じ、財源は２０１９年１０月の消費税増税分の使途変更などで確保する。１９年４月から幼児教育・保育の無償化を一部先行実施し、２０年４月に高等教育を含め全面実施する。幼児教育・保育は、０～２歳児は住民税非課税世帯（年収約２５０万円未満）を対象に無償化する。３～５歳児は、保護者の所得に関係なく認可保育所や幼稚園、認定こども園の利用者は無償化する。認可外施設については、有識者会議を設置して無償化対象などを検討し来年夏までに結論を出す。５歳児については、１９年４月から無償化することを検討している。大学など高等教育は住民税非課税世帯に対し▽国立大は授業料・入学金免除▽私立大は一定上限を設け授業料免除▽給付型奨学金の大幅拡充――を実施。非課税世帯に近い低所得世帯も「非課税世帯に準じた支援を段階的に行う」とした。保育・介護の人材確保に向け、保育士と介護福祉士の賃上げも盛り込んだ。一方、公明党が要望していた私立高校の実質無償化は、２兆円

図 4: 毎日新聞の特徴を表すと考えられる表現のヒートマップの例

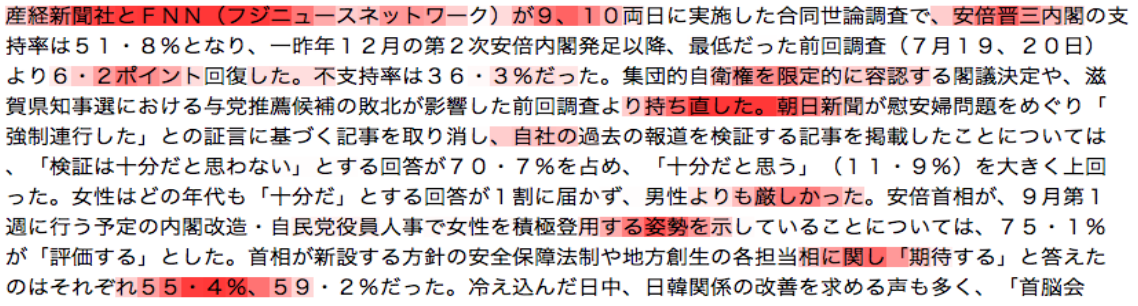
産経新聞社とＦＮＮ（フジニュースネットワーク）が９、１０両日に実施した合同世論調査で、安倍晋三内閣の支持率は５１・８％となり、一昨年１２月の第２次安倍内閣発足以降、最低だった前回調査（７月１９、２０日）より６・２ポイント回復した。不支持率は３６・３％だった。集団的自衛権を限定的に容認する閣議決定や、滋賀県知事選における与党推薦候補の敗北が影響した前回調査より持ち直した。朝日新聞が慰安婦問題をめぐり「強制連行した」との証言に基づく記事を取り消し、自社の過去の報道を検証する記事を掲載したことについては、「検証は十分だと思わない」とする回答が７０・７％を占め、「十分だと思う」（１１・９％）を大きく上回った。女性はどの年代も「十分だ」とする回答が１割に届かず、男性よりも厳しかった。安倍首相が、９月第１週に行う予定の内閣改造・自民党役員人事で女性を積極登用する姿勢を示していることについては、７５・１％が「評価する」とした。首相が新設する方針の安全保障法制や地方創生の各担当相に関し「期待する」と答えたのはそれぞれ５５・４％、５９・２％だった。冷え込んだ日中、日韓関係の改善を求める声も多く、「首脳会

図 5: 産経新聞の特徴を表すと考えられる表現のヒートマップの例